**Assignment**

Welcome to the Nike Engineering Technical Test!

We hope that you find this exercise fun. We want to see your solution to a simple problem with well thought-out and well-structured code. It will be evaluated holistically, so take this as an opportunity to show the breadth of your skills.

The test mainly consists of 2 parts:

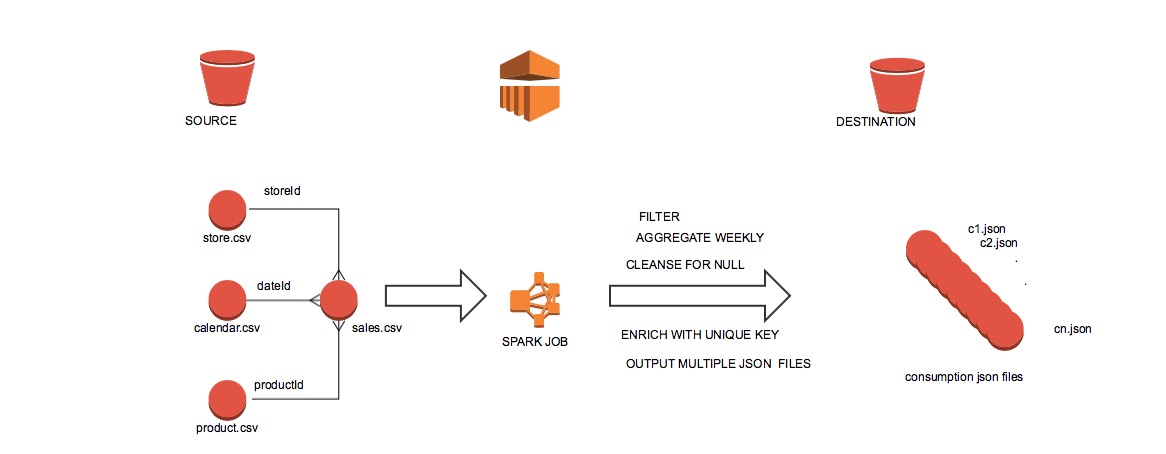
1. Spark Programming task
2. Deployment in AWS EMR (Optional)

**Spark Programming Task**

The aim of this exercise is to create bunch of simple JSON extract files from various csv sources based on the conditions. (weekly aggregator)

**Source files** sales.csv, calendar.csv, store.csv and product.csv

**Destination files format** consumption.json



Sales data contains transactional data with the following columns netSales, salesUnits, dateId, productId and storeId. These fields need to be joined with other fields from other csv’s to get the weekly aggregated value for each combination.

Calendar data contains dateId, week number and year. (Criteria: sales.dateId should be equal to calendar.dateId to get weekNumber and year)

Product data contains productId vs hierarchy. (division, gender, category) (Criteria : sales.productId = product.productId)

Store data contains store vs channel information. (Criteria:sales.storeId = store.storeId)

Output should be bunch of json files which is having the weekly aggregated values of net\_sales and sales\_units. The json files will have unique key. (Refer: consumption.json)

You may use any programming language(**Java/Scala/Python**) with the **Spark**(mandatory). You should use engineering best practices where appropriate. Principles we value include: performance, readability, testability, scalability, simplicity. We should have flexibility to run the spark job based on year and week. You should also aim to achieve a clean separation of concerns between components of your solution; Good luck!

**AWS EMR Deployment (Optional)**

If you have access for the free/paid AWS account, then spin up EMR cluster and demo the spark job during interview. Load the source and destination files to the following s3 buckets

Source Files  
s3://your\_bucket\_name/source/sales.csv  
s3://your\_bucket\_name/source/calendar.csv  
s3://your\_bucket\_name/source/product.csv  
s3://your\_bucket\_name/source/store.csv

Destination Files

s3://your\_bucket\_name/consumption/consumption.json

**The Deliverable**

* A bundled/archived repository showing your commit history or a link to an accessible private repository with your work in (Github Bitbucket). You could fork this repo in git, but any VCS is fine. Git example for sending us a standalone bundle:
* git bundle create <yourname>.bundle --all --branches
* A covering note explaining the technology choices you have made.
* Write a clear and understandable README which explains instructions required to run your solution and tests in a Linux environment.  
  Eg: spark-submit appname arg1 arg2

**Technical questions**

Please answer the following questions

* How long did you spend on the coding test? What would you add to your solution if you had more time? If you didn't spend much time on the coding test then use this as an opportunity to explain what you would add.
* What was the most useful feature that was added to the latest version of your chosen language? Please include a snippet of code that shows how you've used it.
* How would you track down a performance issue in production? Have you ever had to do this the above task with the terabyte of data?